

Cours 1

Statistique descriptive

Ismaël Castillo

École des Ponts, 9 Octobre 2012

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction
QQ-plots

1 Introduction

2 Séries numériques

- Variables discrètes / continues
- Représentation graphique
- Statistiques

3 Deux séries numériques

- Statistiques
- Régression : Introduction
- QQ-plots

1. Introduction

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots

Objectifs

- Définir les quantités statistiques basiques
- Présenter les outils graphiques de la stat. descriptive

On travaillera sur le jeu de données x_1, \dots, x_n sans faire d'hypothèse a priori sur l'existence éventuelle d'un modèle probabiliste sous-jacent

2. Séries numériques

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction
QQ-plots

L'objet de base = les *données*

$$x_1, \dots, x_n$$

Dans ce premier cours, on considère le cas $x_i \in \mathbb{R}$

On parle de *série numérique*.

On distinguera deux types de variables

- les variables *discrètes*
 - ▶ On dit qu'une série numérique correspond à une variable discrète si le nombre de valeurs différentes prises par x_1, \dots, x_n est petit devant n
- les variables *continues*
 - ▶ les autres, typiquement x_1, \dots, x_n correspond à n valeurs distinctes.

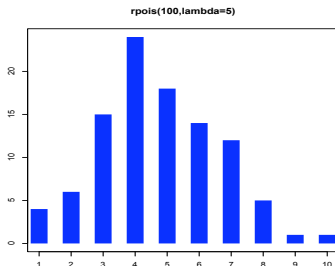
Histogrammes

Stat
descriptive

L'histogramme représente graphiquement le nombre de données par unité/bloc

- Histogramme, cas discret

$$h(x) = \sum_{i=1}^n \mathbf{1}_{x=x_i}$$



Remarque : L'histogramme normalisé est donné par $h(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{x=x_i}$.

Histogrammes

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction
QQ-plots

- Histogramme, cas continu

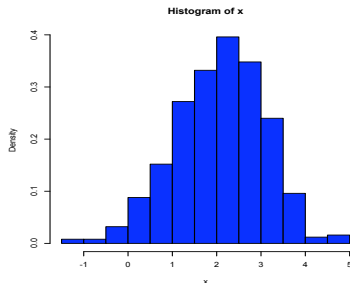
On se donne

- ▶ Un nombre k de classes
- ▶ Une partition de \mathbb{R} en k intervalles I_1, \dots, I_k

$$n_j = \sum_{i=1}^k \mathbf{1}_{x_i \in I_j}$$

Alors

$$h(x) = \frac{1}{n} \frac{n_j}{|I_j|}, \quad \text{si } x \in I_j$$



Histogrammes, choix du nombre de classes

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

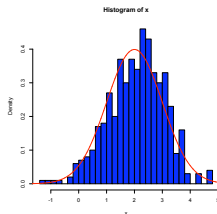
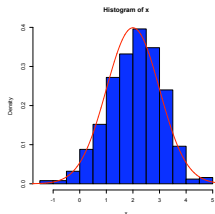
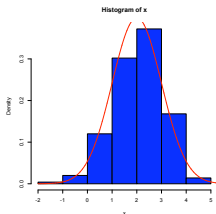
Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots

Les choix de k et de la partition I_1, \dots, I_k sont délicats.
Souvent, on prend

- Une partition uniforme
- On cherche à avoir au moins 5 points par intervalle



Fonction de répartition empirique

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots

Série numérique x_1, \dots, x_n

Definition

La valeur en x de la fonction de répartition empirique associée à (x_1, \dots, x_n) est la proportion d'éléments de la série plus petits que x

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{x_i \leq x}$$

Propriétés

- $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$
- \hat{F}_n est en escalier, croissante
- \hat{F}_n vaut 0 pour $x < \min_i x_i$ et 1 pour $x > \max_i x_i$

Fonction de répartition empirique

Stat
descriptive

Introduction

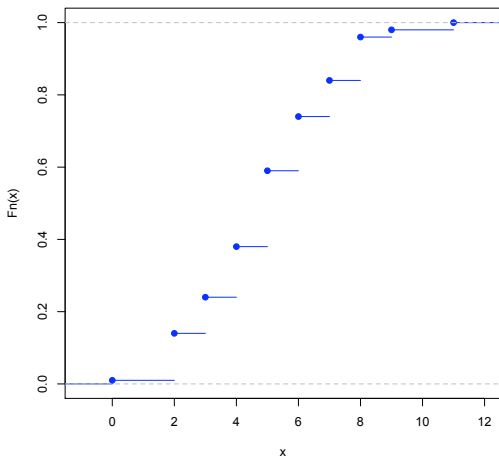
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Exemple 1 : variable discrète
 $n = 100$ x_1, \dots, x_n tirés

selon une loi $\mathcal{P}(5)$

Fonction de répartition empirique

Stat
descriptive

Introduction

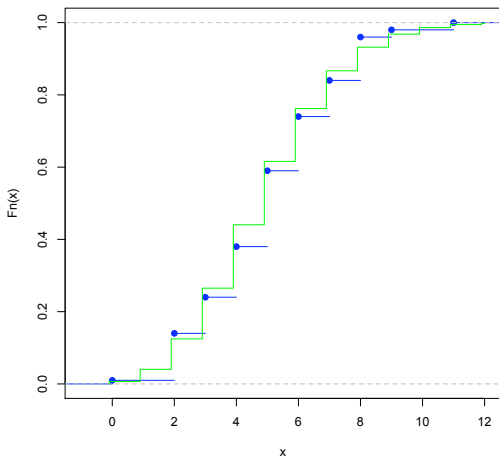
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Exemple 1 : variable discrète
 $n = 100$ x_1, \dots, x_n tirés

selon une loi $\mathcal{P}(5)$

Fonction de répartition empirique

Stat
descriptive

Introduction

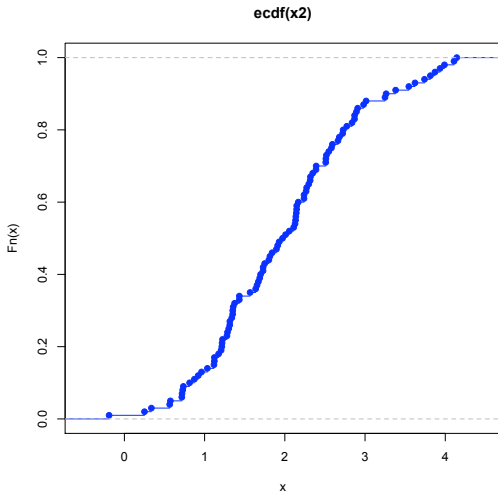
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Exemple 2 : variable continue

$n = 100$

x_1, \dots, x_n tirés
selon une loi $\mathcal{N}(2, 1)$

Fonction de répartition empirique

Stat
descriptive

Introduction

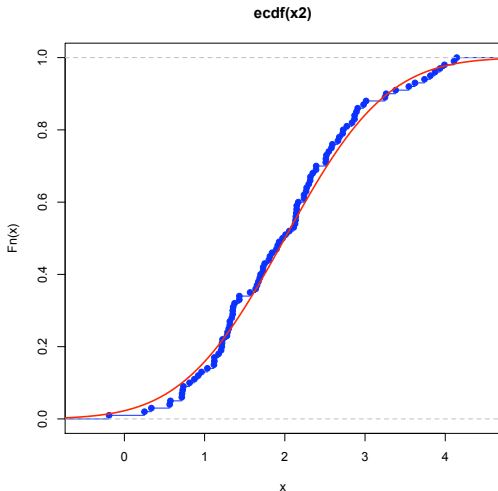
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Exemple 2 : variable continue

$n = 100$

x_1, \dots, x_n tirés
selon une loi $\mathcal{N}(2, 1)$

Statistiques

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction :
QQ-plots

Une **statistique** est une fonction des données, à valeurs dans \mathbb{R}^P

$$S(x_1, \dots, x_n) \in \mathbb{R}^P$$

Exemple $S(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$

Les statistiques sont des *aspects* des données

Idéalement, on cherche un petit nombre de statistiques qui va résumer les données x_1, \dots, x_n . On distingue les

- statistiques de position
- statistiques de dispersion
- statistiques d'ordre (et quantiles)
- ...

Statistiques de position de x_1, \dots, x_n

Stat
descriptive

Moyenne \bar{x}

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Médiane Med_x C'est un nombre m qui sépare les données rangées dans l'ordre en deux ensembles de même taille.

$$x_{(1)} \leq x_{(2)} \leq \dots \mid \dots \leq x_{(n-1)} \leq x_{(n)}$$

Il y a deux cas

- $n = 2p + 1$ impair $x_{(1)} \leq \dots \leq x_{(p)} \leq x_{(p+1)} \leq x_{(p+2)} \leq \dots \leq x_{(2p+1)}$

$$Med_x = x_{(p+1)}$$

- $n = 2p$ pair $x_{(1)} \leq \dots \leq x_{(p)} \leq x_{(p+1)} \leq \dots \leq x_{(2p)}$

$$Med_x = \frac{x_{(p)} + x_{(p+1)}}{2}$$

Remarque. Lorsque n est pair, il y a en général plusieurs nombres qui conviennent. Le choix ci-dessus est habituel.

Exercices et exemples

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Mode *Mode_x* (pour des données discrètes) C'est la valeur la plus fréquente au sein des données.

Exercices et exemples

Stat descriptive

Introduction

Séries numériques

Variables discrètes / continues

Représentation graphique

Statistiques

Deux séries numériques

Statistiques

Régression : Introduction

QQ-plots

Mode *Mode_x* (pour des données discrètes) C'est la valeur la plus fréquente au sein des données.

Exercice. Calculer moyenne, médiane et mode de

$$s = (-2, -1, 0, 5, 8)$$

$$t = (-4, 1, -3, 5, 3, 3, -3, 6)$$

$$x = (1, 1, 2, 3, 3, 3, 3, 9, 20)$$

Exercices et exemples

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Mode *Mode_x* (pour des données discrètes) C'est la valeur la plus fréquente au sein des données.

Exercice. Calculer moyenne, médiane et mode de

$$s = (-2, -1, 0, 5, 8)$$

$$t = (-4, 1, -3, 5, 3, 3, -3, 6)$$

$$x = (1, 1, 2, 3, 3, 3, 3, 9, 20)$$

$$\bar{s} = 2 \quad Med_x = 0 \quad Mode_x = -$$

$$\bar{t} = 1 \quad Med_x = 2 \quad Mode_x = -$$

$$\bar{x} = 5 \quad Med_x = 3 \quad Mode_x = 3$$

Exercices et exemples

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Mode $Mode_x$ (pour des données discrètes) C'est la valeur la plus fréquente au sein des données.

Exercice. Calculer moyenne, médiane et mode de

$$s = (-2, -1, 0, 5, 8)$$

$$t = (-4, 1, -3, 5, 3, 3, -3, 6)$$

$$x = (1, 1, 2, 3, 3, 3, 3, 9, 20)$$

$$\bar{s} = 2 \quad Med_x = 0 \quad Mode_x = -$$

$$\bar{t} = 1 \quad Med_x = 2 \quad Mode_x = -$$

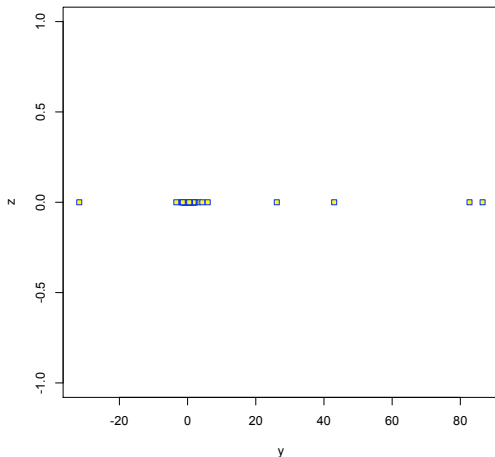
$$\bar{x} = 5 \quad Med_x = 3 \quad Mode_x = 3$$

Illustration phénomène moyenne/médiane

- Salaire net moyen 2008 en France : 2069 euros/mois
- Salaire net médian 2008 en France : 1655 euros/mois

Exemple

- Exemple : moyenne/médiane pour un échantillon de loi de Cauchy



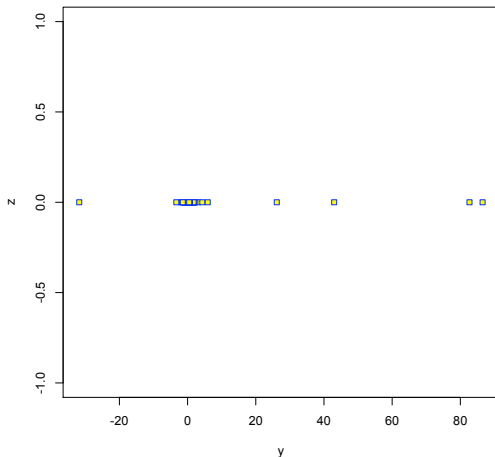
Exemple : Loi de Cauchy

$n = 50$

x_1, \dots, x_n tirés
selon une loi $\mathcal{C}(0, 1)$

Exemple

- Exemple : moyenne/médiane pour un échantillon de loi de Cauchy



Exemple : Loi de Cauchy

$n = 50$

x_1, \dots, x_n tirés
selon une loi $\mathcal{C}(0, 1)$

Moyenne = 4.54

Médiane = 0.27

Statistiques de dispersion de x_1, \dots, x_n

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Variance v_x

$$v_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Écart-type s_x

$$s_x = \sqrt{v_x}$$

Premier quartile Q_1 : médiane des données $< Med_x$

Troisième quartile Q_3 : médiane des données $> Med_x$

Écart inter-quartile : $Q_3 - Q_1$

Remarque : Le deuxième quartile est la médiane des données

Exercices

Stat descriptive

Introduction

Séries numériques

Variables discrètes / continues

Représentation graphique

Statistiques

Deux séries numériques

Statistiques

Régression :

Introduction

QQ-plots

Exercice 1 : Moyenne et médiane d'échantillons

Exercice 2 : Lesquelles des quantités précédentes sont invariantes par permutation des données, par translation des données d'une même quantité μ ? Que deviennent-elles si on multiplie les données par $\lambda > 0$?

Exercice 3 : Distribution exactement symétrique

On dit que x_1, \dots, x_n est (exactement) symétrique par rapport au réel μ si $\forall a > 0$, la fréquence de $\mu + a$ est égale à celle de $\mu - a$.

Calculer la moyenne et la médiane d'une série symétrique par rapport à μ .

Statistiques d'ordre et quantiles de x_1, \dots, x_n

Stat descriptive

Introduction

Séries numériques

Variables discrètes / continues

Représentation graphique

Statistiques

Deux séries numériques

Statistiques

Régression : Introduction

QQ-plots

Il est souvent utile de ranger les données dans l'ordre

$$x_{(1)} = \min_{1 \leq i \leq n} x_i, \quad x_{(n)} = \max_{1 \leq i \leq n} x_i$$

Il existe une permutation $\sigma \in \sigma_n$ telle que

$$x_{\sigma(1)} \leq x_{\sigma(2)} \leq \dots \leq x_{\sigma(n)}$$

On note $x_{(k)} = x_{\sigma(k)}$ la **statistique d'ordre** de rang k .

Statistiques d'ordre et quantiles de x_1, \dots, x_n

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Il est souvent utile de ranger les données dans l'ordre

$$x_{(1)} = \min_{1 \leq i \leq n} x_i, \quad x_{(n)} = \max_{1 \leq i \leq n} x_i$$

Il existe une permutation $\sigma \in \sigma_n$ telle que

$$x_{\sigma(1)} \leq x_{\sigma(2)} \leq \dots \leq x_{\sigma(n)}$$

On note $x_{(k)} = x_{\sigma(k)}$ la **statistique d'ordre** de rang k .

Le **quantile d'ordre α** noté q_α^x est

$$x_{(m)}, \quad \text{avec } m = \lfloor \alpha n \rfloor$$

On peut redéfinir *quartiles* et *médiane* par

$$Q_1 = q_{0.25}^x, \quad \text{Med}_x = q_{0.5}^x, \quad Q_3 = q_{0.75}^x$$

Remarque : peut différer très légèrement de la définition précédente mais pas grave

Box plots (boîtes à moustaches)

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

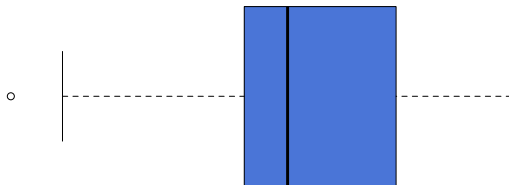
QQ-plots

Un résumé pratique des données x_1, \dots, x_n est donné par

- Med_x , la médiane de l'échantillon
- Q_1, Q_3 , premier et troisième quartiles
- A, B limites en dehors desquelles les données seront considérées comme *aberrantes* ("atypiques", "outliers"). Souvent,

$$A = \min\{x_i : x_i \geq Q_1 - 1.5(Q_3 - Q_1)\}$$

$$B = \max\{x_i : x_i \leq Q_3 + 1.5(Q_3 - Q_1)\}$$



Intérêts

- Résumé des données
- Comparaison d'échantillons

Box plots (boîtes à moustaches)

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

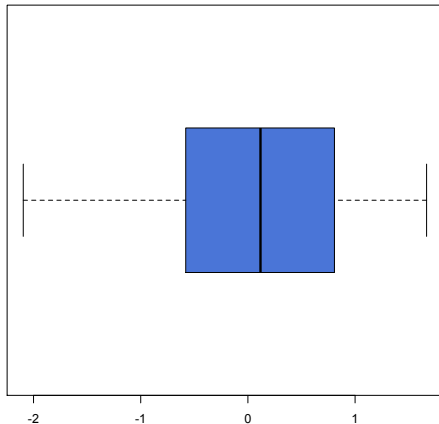
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots



Exemple 1 : loi normale

$n = 50$

x_1, \dots, x_n tirés
selon une loi $\mathcal{N}(0, 1)$

Remarque. Si on prend les quartiles théoriques pour une loi $\mathcal{N}(0, 1)$, la proba pour un tirage x_1 de ne pas être dans $[A, B]$ est 0.7%

Box plots, exemples

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique

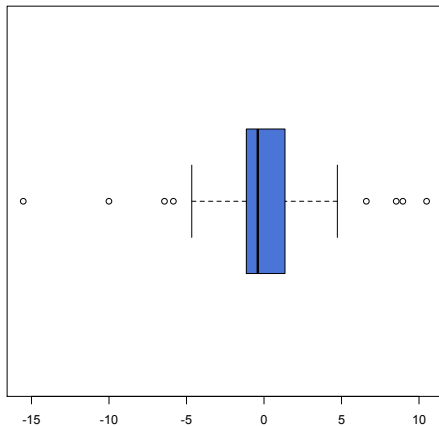
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots



Exemple 1 : loi de Cauchy

$n = 50$

x_1, \dots, x_n tirés
selon une loi $\mathcal{C}(0, 1)$

Comparaison de deux séries numériques

Stat descriptive

Introduction

Séries numériques

Variables discrètes / continues

Représentation graphique Statistiques

Deux séries numériques

Statistiques

Régression : Introduction

QQ-plots

On dispose de deux séries x_1, \dots, x_n et y_1, \dots, y_n qu'on veut comparer

Exemples

- Etude du "lien" éventuel entre x et y
 - ▶ Taille et poids d'un même individu
 - ▶ Température et niveau de pollution à Paris un même jour
- Savoir si x proche d'une distribution théorique donnée (ex. normale)

Covariance et corrélation

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction
QQ-plots

La **covariance** des séries x_1, \dots, x_n et y_1, \dots, y_n notée $s_{x,y}$ est

$$s_{x,y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Covariance et corrélation

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction
QQ-plots

La **covariance** des séries x_1, \dots, x_n et y_1, \dots, y_n notée $s_{x,y}$ est

$$s_{x,y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Le **coefficient de corrélation** linéaire $\rho_{x,y}$ de x_1, \dots, x_n et y_1, \dots, y_n est

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

Covariance et corrélation

Stat descriptive

Introduction

Séries numériques

Variables discrètes / continues

Représentation graphique Statistiques

Deux séries numériques

Statistiques

Régression : Introduction QQ-plots

La **covariance** des séries x_1, \dots, x_n et y_1, \dots, y_n notée $s_{x,y}$ est

$$s_{x,y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Le **coefficient de corrélation** linéaire $\rho_{x,y}$ de x_1, \dots, x_n et y_1, \dots, y_n est

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

Proposition

Pour toutes séries x et y ,

$$-1 \leq \rho_{xy} \leq 1$$

Cas d'égalité : $|\rho_{xy}| = 1$ si et seulement si les séries sont reliées par un relation affine : il existe a, b avec $x_i = ay_i + b$ pour tout $i = 1, \dots, n$.

Covariance et corrélation

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction
QQ-plots

Exercice : Démontrer la Proposition

Nuage de points

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Le **nuage de points** associé aux séries x_1, \dots, x_n et y_1, \dots, y_n est la représentation des points de coordonnées (x_i, y_i) dans le plan.

Parfois, on effectue une transformation préalable des données

Exemple : nuage de points $(\log(x_i), \log(y_i))$

Droite de régression

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Pour un nuage de points $(x_i, y_i)_{i=1, \dots, n}$, notons

- M_i le point de coordonnées (x_i, y_i)
- Δ la droite d'équation $y = ax + b$
- M'_i le point de coordonnées $(x_i, ax_i + b)$
(projection **verticale** de M_i sur la droite Δ)

Droite de régression de Y sur X

C'est la droite qui minimise la quantité

$$\sum_{i=1}^n d(M_i, M'_i)^2,$$

avec $d(M, N)$ distance euclidienne entre les points M et N .

Droite de régression, exemple

Stat
descriptive

Introduction

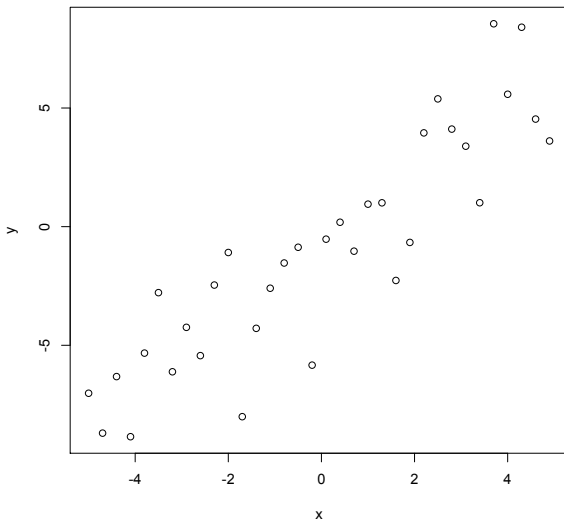
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Droite de régression, exemple

Stat
descriptive

Introduction

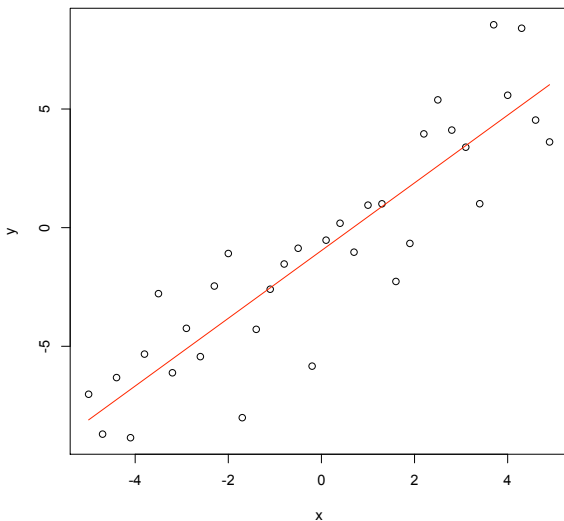
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Droite de régression

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Proposition

L'équation de la droite de régression de Y sur X est donnée par $y = ax + b$, avec

$$a = \frac{s_{xy}}{s_x^2}, \quad b = \bar{y} - a\bar{x}$$

Exercice

- 1 Interpréter géométriquement le coefficient b
- 2 Démontrer la proposition
- 3 Les droites de régression de Y sur X et de X sur Y coïncident-elles ?

Droite de régression, exemple

Stat
descriptive

Introduction

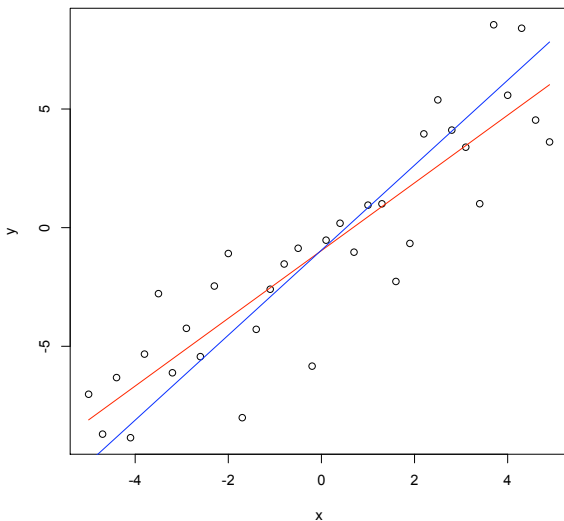
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Droite de régression, exemple

Stat
descriptive

Introduction

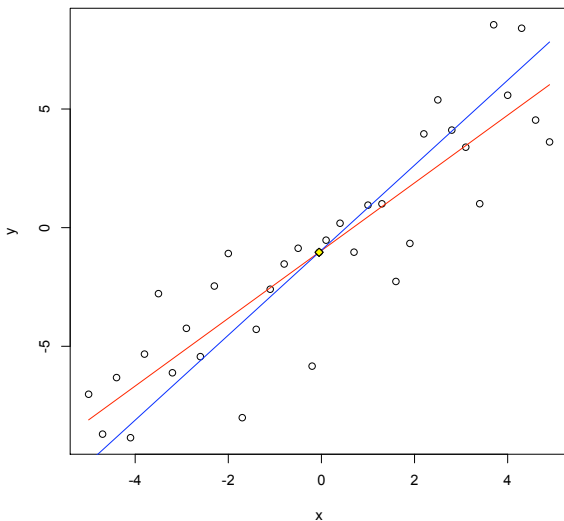
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



QQ-plots

Stat descriptive

Introduction

Séries numériques

Variables discrètes / continues

Représentation graphique Statistiques

Deux séries numériques

Statistiques Régression : Introduction QQ-plots

Premier cas : On cherche à répondre à la question

"Les séries x_1, \dots, x_n et y_1, \dots, y_n suivent-elles la même 'distribution' ?"

Le **QQ-plot** est dans ce cas le nuage de points (q_j^y, q_j^x) , où les q_j^y, q_j^x sont une suite de quantiles de y et x .

Deuxième cas : On cherche à répondre à la question

"La série observée x_1, \dots, x_n se représente-t-elle bien par une certaine loi théorique ?"

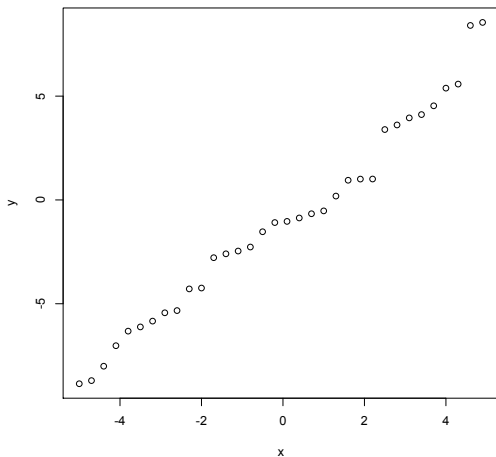
Le **QQ-plot** est dans ce cas le nuage de points (q_j^*, q_j^x) , où les q_j^*, q_j^x sont une suite de quantiles resp. de la loi théorique et des données x .

QQ-plots, exemple, cas 1

Stat
descriptive

Données précédentes "droite de régression"

$$y = ax + b + 2\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1)$$



QQ-plots, exemple, cas 1

Stat
descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

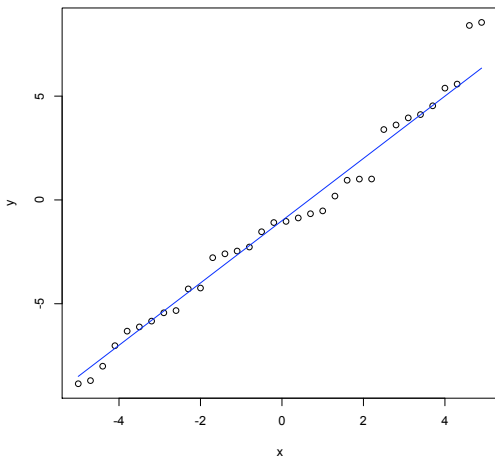
Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots

Données précédentes "droite de régression"

$$y = ax + b + 2\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1)$$



QQ-plots, exemple, cas 2

Stat
descriptive

Introduction

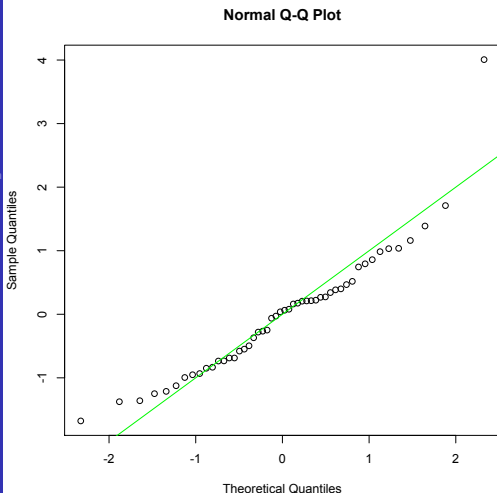
Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques
Régression :
Introduction
QQ-plots



Exemple : loi normale

Échantillon x_1, \dots, x_n
de loi $\mathcal{N}(0, 1)$

QQ-plot

Comparaison à la loi
théorique $\mathcal{N}(0, 1)$

Un dernier exercice

Stat descriptive

Introduction

Séries
numériques

Variables
discrètes /
continues

Représentation
graphique
Statistiques

Deux séries
numériques

Statistiques

Régression :
Introduction

QQ-plots

Exercice : Répartition du PIB/habitant

Faire l'Exercice 1.1 du polycopié